Chapter 3

EXAMINATION OF KNOWN REAL SITES (FROM CHIP-CHIP DATA)

3.1 INTRODUCTION

The "chip-chip" method, introduced in the last few years, seems likely to reveal a vast number of TF binding locations in the human genome. Therefore, an example of these data (Odom et al., 2004) was analysed to see whether it could add substantially to the information produced by analysis of the TransFac data.

Briefly, the chip-chip method is as follows (figure 3.1). First an antibody is prepared which binds to a particular TF. Living cells are treated with formaldehyde; this causes covalent bonds to form between TFs and the DNA they are bound to. The cells are broken open and the DNA sonicated into fragments which, typically, are a few hundred bases long. The antibody is used to precipitate out the TF of interest, with DNA fragments still attached. The covalent links are removed to release the DNA fragments. The latter can be can be analysed in various ways, but for a chip-chip experiment they will be hybridised to a microarray, thus identifying each fragment, provided

it is represented on the microarray.

A disadvantage of chip-chip data is that it does not locate the binding site exactly - typically there is an uncertainty of a few hundred basepairs. A much higher accuracy can be produced by some older techniques, such as promoter mutation.

3.2 METHOD

3.2.1 The choice of TF

The human chip-chip data chosen for study (Odom et al., 2004) was one of the first such studies. Moreover, I had previously studied the TFBSs of HNF1 and HNF4 ((Lockwood and Frayling, 2003), (Lockwood et al., 2003)), which were subjects of the chip-chip study. In addition, (Odom et al., 2004) note their surprise that HNF4 binds to 11-12% of the genes studied (which is roughly equivalent to 1500 genes), they remark that HNF4 "is associated with an unusually large number of promoters", and follow this with additional evidence indicating that the large number is not due to errors. Therefore, focusing on HNF4 could enable a large sample of TFBSs to be covered in one analysis, whereas a different TF would, in most cases, not produce so large a sample.

3.2.2 The PWM and scores

In order to produce a set of results completely independent of TransFac, it was decided not to use any of the TransFac PWMs describing HNF4. Instead, a PWM was complied from a list of 71 HNF4 binding sites that has been published (Ellrott et al., 2002). Two of the sites were reverse complemented because they gave a higher match score in that direction. This produced the frequency matrix in table 3.1.

Figure 3.1: Outline of ChIP method. In drawing 1, the circles represent TFs attached to a length of DNA within a cell. Drawing 2 shows extra bonds, caused by formaldehyde. Drawing 3 shows the cell and DNA broken up by sonication. In drawing 4, the rectangles represent an antibody that attaches to some of the TFs (not all of them, since the antibody is specific to one TF, here referred to as your favourite transcription factor "FTF"). Precipitating the antibody (drawing 5) gives you the DNA fragments bound by FTF (drawing 6).

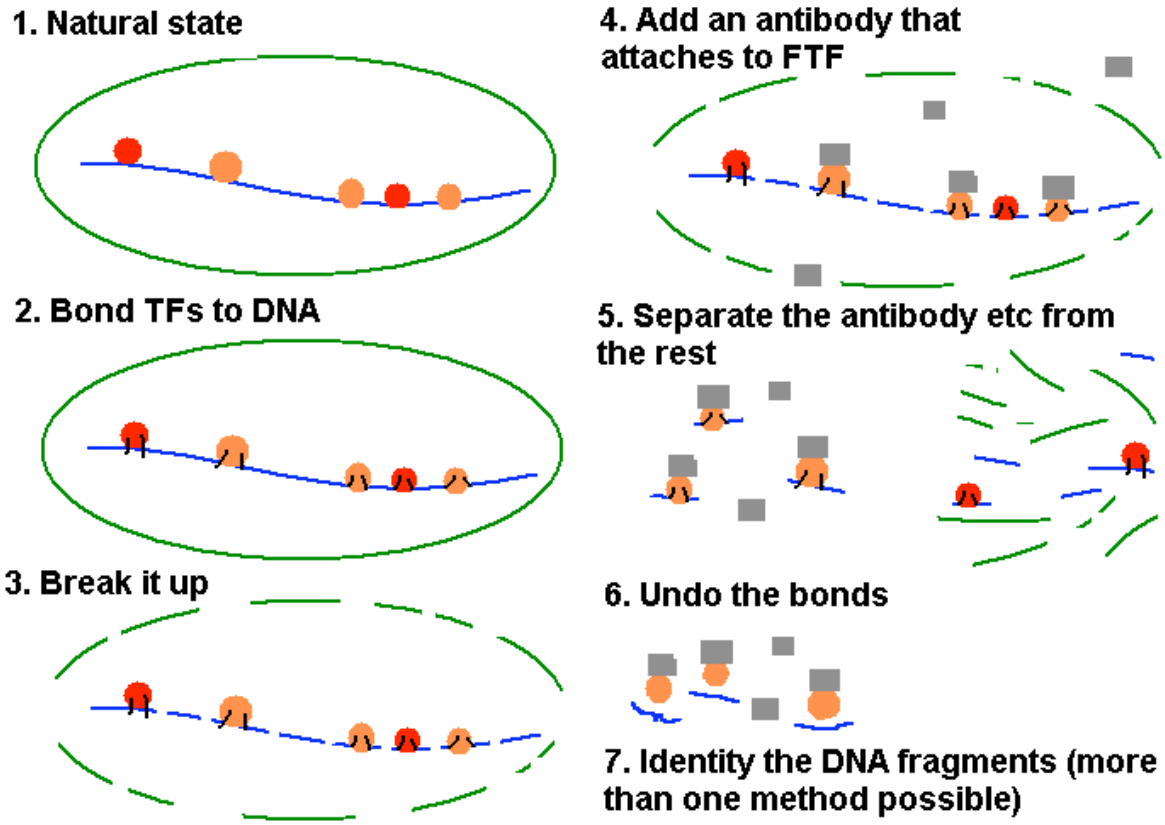


Table 3.1: PWM describing the DNA sequence bound by HNF4
This was based on the list of TFBSs in (Ellrott et al., 2002). Each TFBS is believed to be 13 bases long, hence this table contains 13 rows, each representing one position. For each position, we count the number of times A, C, G and T are present at that position in the list of TFBSs and these form the numbers in the row.

A	C	G	T	Consensus seq
29	7	29	6	AG
3	2	59	7	G
14	4	34	19	AGT
4	23	21	23	CGT
3	53	4	11	C
64	1	2	4	A
57	3	10	1	A
60	2	8	1	A
4	4	62	1	G
6	1	35	29	GT
2	23	12	34	CT
4	53	5	9	C
44	9	10	8	A

HNF4 is a member of the nuclear receptor family of TFs, and binds as a homodimer (Mangelsdorf et al., 1995); and the matrix in table 3.1 contains two approximations to the standard Nuclear Receptor binding sequence (RGKTCA) with a single A between them. That is also true of two other representations of the HNF4 binding sequence, a published sequence logo (Krivan and Wasserman, 2001), and the TransFac matrix M00638 which was at one time was on the "reserve list" of the current project. The matrix shown has an information content of 9.5 bits when examined using information theory (Schneider et al., 1986), which is a technique sometimes used as a measure of specificity. For comparison, the sequence logo has 12 bits and M00638 has 10.9 bits.

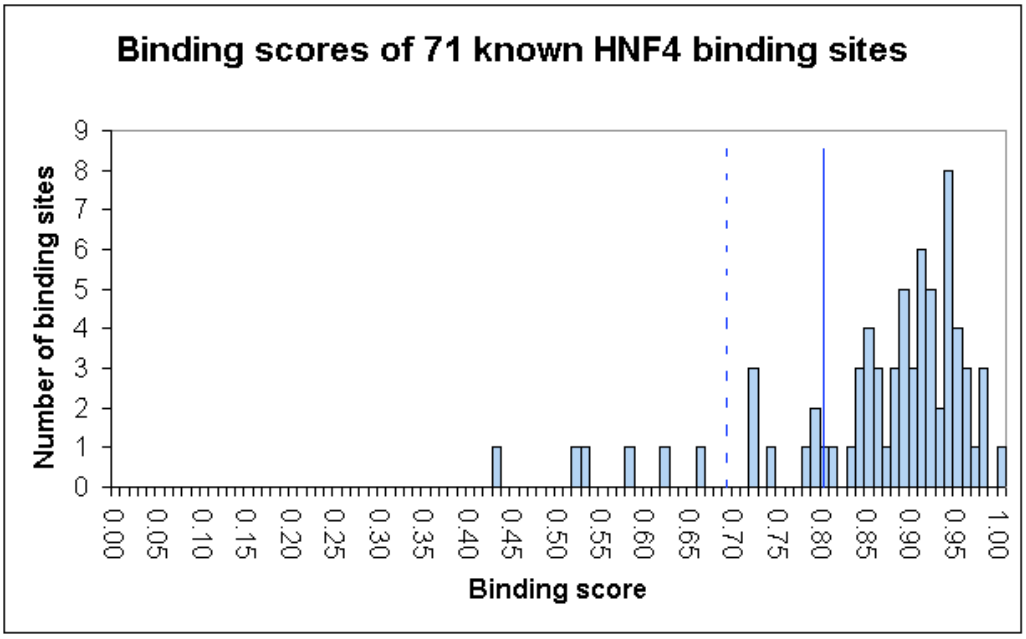
Using this PWM, a binding score was calculated for each of the 71 binding sites; a histogram of these scores is shown in figure 3.2. In an earlier chapter (page 94), it was described how a distribution of this kind was used to set the "threshold score", that is, the score which it was thought would generally be exceeded by HNF4 binding sites. The same procedure was adopted here.

However, the setting of the threshold score and standard deviation was complicated by the presence of a small number of very low scoring TFBSs, which can be clearly seen in the figure 3.2 histogram. Is there an argument for leaving these out? If one has 71 scores that follow the Normal (or Gaussian) distribution, one would not expect any scores to be more than 2.5 standard deviations below the average (because an individual score has a 0.006 probability of being below this value, so a sample of 71 scores is unlikely to produce a single score like this). But, in fact, there were three very low scores that were below this value. Leaving these three TFBSs out and recalculating the standard deviation, then repeating this a number of times, the sample stabilises with 10 TFBSs left out and the standard deviation being just under half its original value. Consequently, the threshold score (which depends on this distribution) is set to a value indicated by the solid blue line.

The main analysis used the standard deviation which has been estimated after leaving out the 10 low scoring TFBSs. The argument for doing so is this: the main use of the standard deviation is in deciding the importance of a

difference between two scores of which one (at least) is above the threshold. Therefore, it seems sensible to take all the scores which form the "bell" around the threshold and use those to estimate the standard deviation. The very low scoring TFBSs, which form a "tail" distribution that stretches far below the threshold, do not appear to be relevant.

Figure 3.2: This histogram shows the binding scores for the 71 known HNF4 binding sites. Note that there are a small number of sites which have extremely low scores; should they be included when calculating the standard deviation, 20th percentile, and the threshold score which is derived from them? If all binding sites are included, then the threshold will be as shown by the dashed blue line. However, if the lowest scoring sites are omitted, the threshold will be as shown by the solid blue line. The latter was adopted. Broadly speaking, it is assumed that a sequence with a binding score below the threshold score is not a TFBS (but for a more detailed description of how the threshold is calculated and used, see page 94).



However, the matrix was still based on all 71 TFBSs, including the very low scoring TFBSs, as the latter presumably contain data about weak but permissible DNA contacts that might also be used in higher scoring TFBSs. Possibly these low-scoring TFBSs are the reason why the matrix has a lower information content than other published representations of HNF4 binding.

The list of 71 TFBSs used nonstandard gene names. The species containing each TFBS was not given in the publication, but could be obtained from an author's website (http://www.sladeklab.ucr.edu/hnf43.pdf). Some of these were given a standard gene symbol, with GeneLynx (Lenhard et al., 2001) being useful to resolve some of the more obscure cases, but this was not always successful. Therefore, whilst all 71 were used to form the PWM, not all could be used in other analyses.

3.2.3 The initial list of genes regulated

The authors (Odom et al., 2004) have made the detailed results of that study available on their website, at

http://jura.wi.mit.edu/young_public/pancregulators/HNF_data_v2.xls

From these data, the main data extracted for each entry were the gene name, RefSeq ID, "HNF4alpha pvalue HEP" and "HNF4alpha pvalue ratio HEP". Thus, the hepatocyte data were used; the data from pancreatic islets were not used. The reason for ignoring the pancreatic data was that, from their supplementary table S14, the success rate for finding the exact TFBS for HNF4 was higher in hepatocytes than in pancreatic islets.

Their supplementary material states that "Genes were included in the set of 'bound' genes if the binding P-value in the error model was < 0.001 or enrichment was at least 2-fold in the immunoprecipitation"; so only if a gene met this criterion would it be included in the analysis.

3.2.4 The size of the upstream search region

Chip-chip data only locates a TFBS to an accuracy of a few hundred bases, so the exact location had to be determined by searching that length of DNA using the PWM that describes HNF4 binding sites (see table 3.1). Searching a long length of DNA increases the risk of obtaining a match that is incorrect, even though it has a high binding score. For that reason, it is desirable to keep the search region as small as possible.

The group that produced the chip-chip data ((Odom et al., 2004) , supplementary material) used a microarray with 1kb DNA, representing −750 to +250 relative to the transcription start site. A TFBS slightly outside this region might still be detected because of the length of the DNA fragments produced by sonication. Presumably, this is why they searched −1000 to +500 for matches to TransFac HNF4 matrix M00134. However, that may be too large a search region.

An earlier study of mine (Lockwood and Frayling, 2003) searched for HNF1 sites as far upstream as −2000, but concluded that this may have been too large, since searching up to −500 would still have found 7 out of the 8 sites that were identified. Thus for HNF1, this is justification for a search to go no further upstream than −500. It is not clear whether this also applies to HNF4, but it might, since both factors are known to act synergistically (Miura and Tanaka, 1993) and both have a role in directing liver-specific transcription (Krivan and Wasserman, 2001). Also, it has been claimed that the binding sites for HNF4 are nearly always within 600 bases of the TSS (Levitsky et al., 2002). Therefore a search from −500 to 0 only was used. This is smaller than the search region that the original authors used on the same data, as described in the previous paragraph.

3.2.5 The analysis of whether the TFBS was conserved

The analysis followed the procedure described in the previous chapter for TransFac TFBSs.

In computational terms, locating the TFBS using a 500bp search was not new, as the TransFac data had required the same kind of search (as described on page 87); this had been required because TransFac site sequences often represent a slightly incorrect location. In statistical terms, however, the 500bp search was a serious change since it was so much larger than was usual with TransFac data (where as few as 15 possible positions might be searched). This larger search greatly increased the risk of finding a high-scoring but incorrect match to the PWM, therefore making it more important to have a

sensible procedure for rejecting incorrect matches of this type.

The procedure was as follows. For each particular gene, a “p-value” was estimated by searching the DNA 500bp upstream using 100 shuffled matrices, finding the highest scoring match for each shuffled matrix, and finding how many shuffled matrices produced a binding-score higher (or as high) than the real matrix did. Therefore the p-value could take any of 101 possible values; these would be equally likely to occur if the HNF4 site was not in the search region, and if the shuffled matrices were an accurate model of how scores are distributed when there is no TFBS. Again, this reused a procedure that had already been used, since the 100-shuffled-matrix-search had aleady been used with the TransFac data, as described by the flowchart in figure 2.3. However, the 10% criteria specified in that flowchart was chosen for TransFac data and might not be appropriate to the chip-chip data, especially since (as noted in the previous paragraph) the chip-chip data created a much higher risk of getting a high-scoring but incorrect match. The 10% criterion is equivalent to accepting a gene if its p-value is 0.1 or less, but since that criterion may not be the ideal choice for chip-chip data, the results section will consider what sensible p-value criterion should be adopted.

3.2.6 "Second-best matches"

The "second-best match" procedure was originally used with the TransFac data (page 91), but the nature of the chip-chip data made the value of a particular parameter much more critical. To recap, this procedure was used when the search of the DNA (which was believed to contain a TFBS) found two good matches, and the two binding-scores were very similar. It would be unsafe to assume that the highest-scoring match was definitely the true TFBS, and so both possible TFBSs would be analysed. If the two analyses came to inconsistent conclusions then the overall conclusion for the TFBS would be "uncertain". Similarly, if their scores were close to the best match, then the 3rd-best site would be analysed, etc, up to but not beyond the 7th-best site.

A critical parameter, called "$scoreMargin", defined what difference indicated that binding-scores were close enough for the second-best match to require analysis. For the TransFac data, it was set to twice the standard deviation of the scores of real TFBSs (using the real TFBSs from which the PWM was compiled). This, it was assumed, would correspond to a sufficiently large difference in binding affinity that the second-best match was very unlikely to be the true TFBS, and could be ignored if its score was that far below the score of the best match.

For example, applying this to the HNF4 chip-chip data, if the search of 500bp reveals a best-match with a binding score of 0.95, and a second-best match with a binding score of 0.94, then one could plausibly argue that either match could be the real TFBS, since the difference between them is so small (looking at figure 3.2 suggests the difference between 0.95 and 0.94 is trivial). Conversely, suppose the best-match had a binding score of 0.97, and the second-best match had a binding score of 0.82, then it seems far more likely that the best-match is the true TFBS, and rather unlikely that the second-best match is the true TFBS, so it is justifiable to ignore the latter (looking at figure 3.2 suggests that the difference between scores of 0.97 and 0.82 is an important difference, as the former will be a TFBS that binds more strongly than average, whereas the latter will bind rather weaker than the average TFBS). "$scoreMargin" represents a difference that is just large enough to justify ignoring the second-best match.

To choose a sensible value of "$scoreMargin", note that one study found that high-scoring TFBSs tend to have a noticeably higher binding affinity than low-scoring TFBSs (using *in vitro* measurements) (Tronche et al., 1997). Thus, for a histogram like figure 3.2, the TFBSs on the right side of the peak are likely to have a noticeably higher binding affinity than the TFBSs on the left side of the peak. Thus a noticeable difference in binding affinity will be represented by a difference in score that is the same size as the "spread" of score within the peak on figure 3.2. This suggests using the standard deviation of this peak as being equivalent to a substantial change in binding affinity. (However, it is unclear what exact value to choose - one standard deviation

or two standard deviations).

For the initial analysis of the chip-chip data, the value of $scoreMargin was set on the same basis as had been used for the TransFac data - that is, $scoreMargin was set to twice the standard deviation of the distribution shown in figure 3.2 (leaving out the outliers). However, now the search was of a region as large as 500bp, it resulted in the second-best analysis being triggered very frequently (76% of cases, and in 56% of cases the third-best site was analysed). It only required one of these other sites to generate a conclusion inconsistent with the best site for the TFBS to be marked "uncertain". This could cause a large portion of the data to be marked "uncertain". To avoid this, and since the original setting of $scoreMargin was somewhat arbitrary, it was decided to try some analyses with $scoreMargin set to one standard deviation only, though some analyses used $scoreMargin set to two standard deviations.

3.2.7 Fictional TFBSs

Fictional TFBSs were generated using 1169 shuffled matrices. As with the TransFac-type analysis, fictional TFBSs were selected from a stretch of DNA 50 bases either side of the DNA that was thought to be the real TFBS. The "real" TFBS was taken to be the best match within the search region (from 500 bases upstream of the start of the gene), and a real TFBS would only be used for this purpose if it had a sufficiently high score (to meet the $p<0.1$ criterion).

3.3 RESULTS

3.3.1 The proportion of TFBSs that are successfully located and the choice of p-value

For each gene regulated by HNF-4 (as shown by the chip-chip data), a p-value was estimated, based on how good a match to the PWM was found after searching the 500bp upstream. Figure 3.3 shows a histogram of the p-values for 994 genes. If a gene had an excellent match to the PWM (in the upstream 500bp), then it would have a very low p-value, and hence if there were many genes with good matches, the histogram would show a peak bunched towards the left side of the graph, which it does in fact have. Thus the peak on the left suggests many HNF-4 TFBSs have been successfully located.

However, suppose the HNF-4 binding site for a particular gene was not successfully located (eg, because the true TFBS was outside the 500bp search region); then, by the definition of p-value, one would expect that gene to have a 1% chance of getting a p-value between 0 and 0.01; a 1% chance of getting a p-value between 0.01 and 0.02; a 1% chance of getting a p-value between 0.02 and 0.03; etc. Thus if the search failed to locate the true TFBS for every gene, the histogram would have a (roughly) flat top, as the genes would be evenly distributed over all the p-values. The right side of figure 3.3 is indeed like that. The right side of the histogram is made up entirely from genes with poor p-values (which indicate that the matches to the PWM could be chance matches rather than true HNF-4 binding sites), and does have a roughly flat top, as indicated by the red dashed line. Since the bars on the right side of the histogram ($p>0.5$) have an average height of 6.9, the red dashed line has been placed at that height.

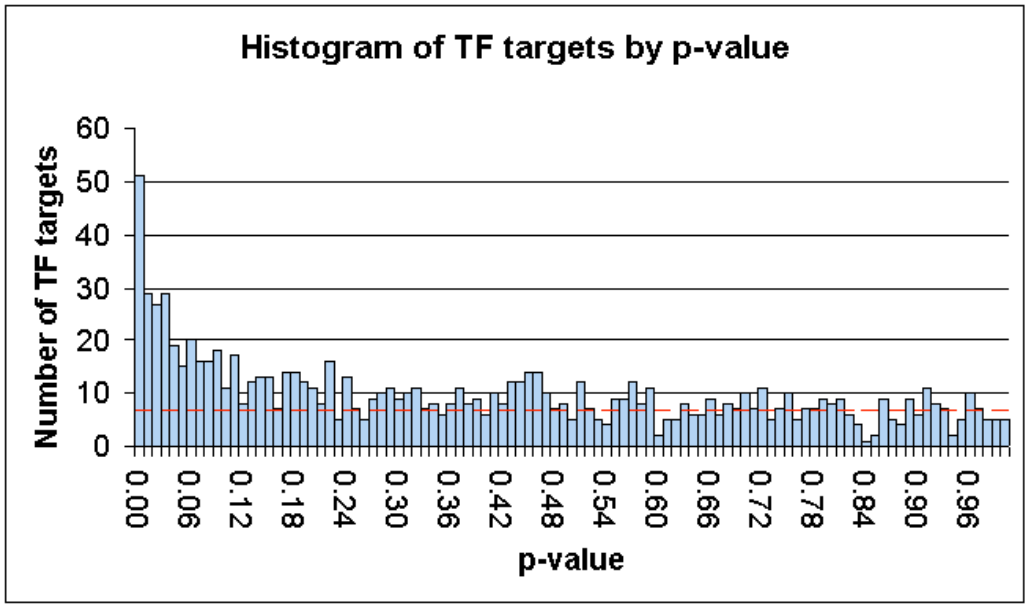
Thus the right side of figure 3.3 can plausibly be interpreted as being largely caused by genes for which the PWM search did not find the true TFBS, but found a chance match instead; and the red dashed line indicates the frequency with which such matches are expected to occur.

Turning attention back to the left side of the histogram, it is clear that many bars are much higher than the red dashed line, and the red dashed line indicates the number of matches expected to occur by chance. For instance, the highest bar contains 51 TFBSs, whereas 6.9 are expected to be produced from searches where the correct TFBS was not located. Thus the difference between these two figures (44.1) is an estimate of the number of correctly located TFBS represented by this particular bar of the histogram. By producing a similar estimate for each bar in the histogram and adding, it suggests a grand total of 297 TFBSs were correctly located. Hence, of the 994 TFBSs, the remaining 697 TFBSs are likely to have been incorrectly located by the 500bp search process.

It is evident that there is serious contamination from TFBSs that were not correctly located. As just noted, these are expected to make up 6.9 of the 51 TFBSs that make up the highest bar of the histogram (note that this bar represents p-values less than 0.01). Hence if those were used as a sample, the contamination rate would be that 14% of TFBSs that are incorrectly located, and the sample would only include 51 TFBSs. But the purpose of using the chip-chip data was to produce a large sample and greatly add to the data available from TransFac, so a sample of 51 TFBSs is far too small - and whilst we can increase the number of TFBSs in the sample by using a higher p-value, this makes the contamination problem worse. For example, if we accept all TFBSs with a p-value less than 0.05 (represented by the 5 bars on the left of the histogram), then 155 TFBSs will be obtained, but it is estimated that 22% of these have been incorrectly located by the 500bp search. If we accept all TFBSs with a p-value of 0.10 or less, 251 TFBSs will be obtained, but it is estimated that 30% will be incorrectly contaminated.

Thus, any sample from this data will be contaminated by about 20-30% of the TFBSs having the wrong location assigned by the 500bp-search. The exact contamination rate can be altered a little depending on the exact p-value criterion chosen, but that will not make a drastic difference.

Figure 3.3: The chip-chip data (Odom et al., 2004) identified a large number of genes as being targets of the HNF-4 transcription factor. Each gene was given a p-value, by searching 500bp upstream of the gene for the best match to the HNF-4 PWM; the p-value is an estimate of the probability of obtaining a best match that good by chance. If none of these 500bp regions really contained an HNF-4 binding site, then one would expect the p-values to be evenly distributed; that is, this histogram would be shaped roughly like a rectangle, with a flat top, as suggested by the dashed red line. However, the actual histogram shown here has a peak on the left. This shows that very low p-values are far more frequent than one would expect by chance, suggesting that HNF-4 binding sites do indeed exist in most of the regions which have very low p-values.



3.3.2 Conservation of DNA surrounding conserved and diverged TFBSs

The most interesting finding from the TransFac data was that DNA flanking conserved TFBSs is more highly conserved than the DNA flanking diverged TFBSs. That difference was larger for real TFBSs than it was for fictional TFBSs. A similar analysis of the HNF4 chip-chip data was therefore undertaken.

As the table 3.2 shows, the finding from the TransFac data was not completely replicated. For the HNF4 chip-chip data, there was almost no difference between real and fictional TFBSs. This is particularly difficult to explain because the clearest difference between chip-chip and TransFac results is for the two sets of *fictional* data - for the real TFBSs, chip-chip and TransFac give similar results.

Table 3.2: Neighbourhood conservation by data source
Conservation of DNA flanking each TFBS: the difference between conserved and diverged TFBSs, by source of data. The measure of conservation is the %identity of DNA flanking the TFBS, in a two-species comparison.

	Real TFBSs	Fictional TFBSs	$scoreMargin
HNF4 chipchip data (run H3/H3)	$8.4\% \pm 2.1\%$	$7.4\% \pm 1.2\%$	2*s.e.
TransFac data (run J)	$9.3\% \pm 1.9\%$	$5.6\% \pm 0.9\%$	2*s.e.,1*s.e.
TransFac data, human source (run J)	$9.1\% \pm 2.1\%$	$6.4\% \pm 1.0\%$	2*s.e.
HNF4 TFBSs from (Ellrott et al., 2002)	Too few TFBSs	$2.8\% \pm 2.3\%$	

"DNA surrounding each TFBS" means 50 bases on each side, not including the TFBS; the difference was calculated using the stratified-by-species method

To examine the reason for the difference between the TransFac and chip-chip results, a number of other runs were done, and these are also shown in table 3.2.

The chip-chip data came from an experiment on human cells, whereas the TransFac data came from experiments on human, mouse and rat. To see what

effect (if any) this difference would have, a human-only subset of the TransFac data was analysed. This is shown in the table as “TransFac data, human source" and it can be seen that the size of effect for fictional TFBSs, 6.4%, is about midway between the corresponding values for original TransFac (5.6%) and for chip-chip (7.4%) data. Given the size of the errors, this does not lead to any firm conclusion.

The chip-chip data used related only to the TF HNF4, whereas the TransFac data related to a number of different TFs; thus a difference between the two might have been caused by HNF4 being different from other TFs. To investigate this, an analysis was done using the TFBSs which had been used to compile the HNF4 PWM (shown in table 3.1). Although there were 71 TFBSs on that list, many could not be used because of nonstandard gene names, or failure to find the given binding sequence in the DNA sequence upstream of the gene. Consequently only two were identified as examples of diverged TFBSs, and so no result has been entered into the "Real TFBS" column of the table. However, the fictional TFBSs generated from this sample produced many more examples (78 diverged TFBSs) and for these, the size of the effect is 2.8%. If this was clearly larger than the corresponding value from TransFac data (5.6%), it would be evidence that HNF4 behaved differently to other TFs; but of course it is not larger. So this analysis did not produce any evidence suggesting that HNF4 fictional TFBSs are different from the fictional TFBSs of other TFs.

Neither of the above provide a positive explanation for the difference between the TransFac and chip-chip results. Another possibility, to be considered later, is that it was caused by the different method of analysis that had to be adopted with the chip-chip data, in particular the 500-bp search. Whether the latter was important was not always clear; for fictional chip-chip data, a preliminary run where the 500-bp search was not used gave similar results to a later run where the 500-bp search was used. However, a detailed investigation (to be described in section 3.4) suggested the 500-bp search was important.

3.4 USING TRANSFAC DATA TO ESTIMATE ERRORS FROM CHIPCHIP DATA

3.4.1 Method

A problem with chip-chip data is that the position of a binding site is only found to an accuracy of a few hundred bases. In this project, the exact position was determined by searching that region for the best match to the PWM. How did this affect the final results, compared with what would be obtained if the exact location were available?

This was addressed by analysing the TransFac data as if it was chip-chip data. The TransFac information on the coordinates and DNA sequence of the TFBS were ignored; the exact location of each TFBS was determined instead by searching for the best match within 500 bases upstream of the gene. This, of course, introduced a risk that the incorrect location would be chosen, just as can happen when true chip-chip data is analysed. TransFac TFBSs that have been used in this way will be called "pseudo chip-chip data".

After this work had been completed, additional data became available, due to more data being added to the the TransFac database. Thus the results described in this section are based on the earlier set of TransFac data, not the same set of data used to generate results in other chapters of this thesis. Each time the analysis program was run, the results from that run were given an identity code, such as "Run J". These will be referred to here, so it is clear when two results are derived from the same dataset. "Run J" was based on TransFac version 7.4.

The "Run J" fictional TFBSs were converted to pseudo chip-chip fictional TFBSs by producing a datafile with the position and DNA sequence deleted; the results from this will be referred to as Run M pseudo chip-chip fictional TFBSs.

3.4.2 Results

From table 3.3, there does appear to be a noticeable difference caused by converting the run J fictional data to pseudo-chip-chip fictional data; the difference in %id neighbourhood increases from 5.5% to 8.1%-10.2%. The 8.1%-10.2% range is close to the 7.4% obtained in the analysis of HNF4 chip-chip data (see table 3.2), which suggests that the pseudo chip-chip fictional data is successfully modelling the problem we are examining.

At one time it was thought that the exact value of $scoreMargin might be critical (based on evidence from a small sample), and therefore table 3.3 shows run M results for different values of $scoreMargin. In fact, as just noted, all the values exceed the 5.5% for the run J fictional TFBSs, so a explanation purely in terms of $scoreMargin does not look possible.

Table 3.3: Pseudo chip-chip: %id of neighbourhood: the difference between conserved TFBSs and nonconserved TFBSs, for different analyses (all based on TransFac data, which, however, was altered to resemble chip-chip data in the cases marked as "pseudo chip-chip")

Type of data & analysis	$scoreMargin set to	difference in %id of neighbourhood	
		mean	error
RunJ real TFBSs	2*s.d.	9.3%	1.9%
RunJ fictional TFBSs	2*s.d.	5.5%	0.9%
RunM pseudo chipchip fictional TFBSs	2.5*s.d.	8.1%	3.0%
RunM pseudo chipchip fictional TFBSs	2*s.d.	8.9%	1.2%
RunM pseudo chipchip fictional TFBSs	1.5*s.d.	10.2%	0.8%
RunM pseudo chipchip fictional TFBSs	1*s.d.	10.0%	1.1%
RunJ fictional TFBSs, for which a full analysis had been made during run M	2*s.d.	8.1%	1.8%

The %id of neighbourhoods are all species-stratified

Although the pseudo chip-chip data seems to have successfully modelled the problem, for a fuller understanding of the problem further investigations were necessary. There were some TFBSs for which an analysis against comparison species was not attempted, generally because a problem in identifying the known TFBS. In some cases, such a problem occurred during run M even

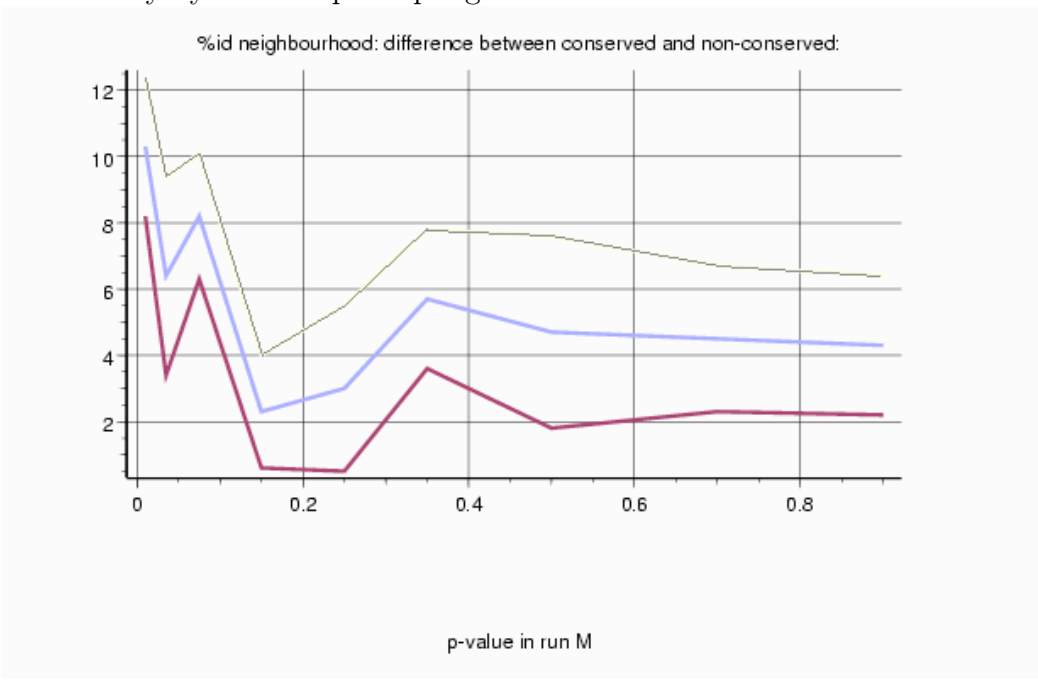
though it had not during run J for the same TFBS. To examine the effect of this, these TFBSs were excluded from run J (fictional), and the remaining subset of TFBSs analysed (as ordinary fictional TFBSs, that is, not as pseudo chip-chip data). As the bottom entry of table 3.3 shows, excluding these TFBSs caused the difference in %id neighbourhood to increase from 5.5% to 8.1%. This suggests that, when TFBSs are excluded because of problems with the TFBS, this alone will have a serious effect.

This suggests that the key to the problem was the exclusion of TFBSs, but why did this happen? Many of the TFBSs were excluded because of their p-value (exclusion occurring if $p > 0.1$). As already noted (section 3.2.5), because of the 500-bp search, such exclusions were far more common with chip-chip data than when analysing TransFac-style data. The reason for this can be illustrated by an example: a TFBS with a binding-score of 0.83 would be given a p-value of 0.29 if it was found by searching 500bp, yet the same TFBS (binding-score=0.83) would be given a p-value of just 0.01 if it was found by searching 20bp (as typically happens when analysing TransFac data). The precise p-value would depend on the PWM and GC content, but the p-values just given in this example are a typical illustration. The underlying reason is that a TFBS with a binding score of 0.83 is much more likely to be found by chance when 500bp are searched than when, say, 20bp are searched. Hence the p-value of a particular TFBS must depend heavily on the size of the search region.

It was therefore decided to examine possible evidence that excluding TFBSs with $p > 0.1$ would affect the results from the run M pseudo chip-chip fictional data. To do this, subsets of the run J fictional TFBSs were produced; each subset contained TFBSs whose run M p-value was in a particular range (eg, $0.02 \leq p < 0.05$). Thus, the p-value came from pseudo chip-chip analysis, but in all other respects the analysis did not involve any kind of chip-chip data; this was the means of isolating the effect of p-values from other features of the chip-chip analyses. Figure 3.4 shows results from this data and it can be seen that points for which $p \leq 0.1$ do appear to differ from $p > 0.1$ points. It is therefore plausible to suggest that this would affect results that

involved the difference in %id of neighbourhood sequence between conserved and diverged TFBSs. A higher value would be obtained if $p > 0.1$ TFBSs were excluded, as happened in the pseudo chip-chip analysis (run M), than if they could be included, as in run J.

Figure 3.4: %id of neighbourhood: variations with p-value. Each point shows (on the y-axis) how much %id of the neighbourhood sequence differs between conserved and diverged fictional TFBSs. Each point represents a subset of the run J data, consisting of the fictional TFBSs for which the p-value was within a particular range, the p-value being from the run M pseudo chip-chip data. The middle line shows the observed values, and the upper and lower lines are plus or minus one standard error. Each point had its error estimated individually by bootstrap sampling.



3.5 SUMMARY AND DISCUSSION

The work reported in this chapter failed to reproduce the key result from the earlier TransFac work, concerning the extent to which DNA surrounding the TFBS is conserved. Even more alarmingly, the "fictional TFBS" result from

the chip-chip data was close to the "real TFBS" result from the TransFac data (table 3.2). Was this merely coincidence, or was it evidence that the "real TFBS (TransFac)" result was seriously affected by some problem that had been overlooked when that data were analysed, but that had been reproduced by the "fictional TFBS" analysis of the chip-chip data? The latter possibility would be very serious, undermining the key result from the TransFac-based work. Therefore a lengthy attempt was made to understand why the chip-chip data resulted in a "fictional TFBS" effect larger than that from the TransFac data, despite the number of details that had to be considered.

One possible reason was that the effect was peculiar to HNF4, the TF on which all the chip-chip data was based; however, analysis of HNF4 TFBSs from a different source did not support this idea. Another possible reason was that the chip-chip data came entirely from human cells, and not rodent cells also; investigation failed to find evidence in favour of this or to rule it out.

A more plausible reason was that the chip-chip data only locates a TFBS with an accuracy of a few hundred basepairs, and so the exact location has to be found by using a PWM to search a region of (in this case) 500bp. Evidence for this came from analysing TransFac data as "pseudo chip-chip data", which produced a "fictional TFBS" result closer to that obtained from fictional chip-chip data. Therefore, it is plausible to think that the chip-chip data might have given results similar to the TransFac data if the 500bp search could have been eliminated, or at least drastically reduced in size. But this would have required chip-chip data that gave a much more accurate location that the data actually used.

Problems related to the 500-bp search were also noted on page 148, where it was estimated that 20-30% of TFBSs will have been incorrectly located by the search.

Future chip-chip data may reduce the size of the search region required. The data used here came from a printed microarray, where each DNA molecule on the microarray was 1kb long, and that would restrict the accuracy of location. But much shorter lengths are possible, especially when the DNA

is synthesised on the microarray rather than printed. Improved data are also likely to come from a new ChIP technique which eliminates the microarray, instead using high-throughput DNA sequencing; a trial has shown this can locate large numbers of TFBSs with an accuracy of less than 50bp (Johnson et al., 2007). Thus future ChIP data may well give more accurate locations and reduce the size of the search region required.

Declaration: no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning, except for the reuse of some Perl modules (altered as required), originally written during the candidate's Master in BioInformatics course (University of Exeter)